Reply to T. Schneider's comment on "Spatio-temporal filling of missing points in geophysical data sets"

D. Kondrashov, M. Ghil

Nonlin. Processes Geophys., 14, 3–4, 2007
www.nonlin-processes-geophys.net/14/3/2007/

Nonlinear Processes in Geophysics

Reply to T. Schneider's comment on "Spatio-temporal filling of missing points in geophysical data sets"

D. Kondrashov[1] **and M. Ghil**[1,*]

[1]University of California, Los Angeles, CA, USA
[*]Ecole Normale Supérieure, Paris, France

Received: 9 August 2006 – Revised: 21 December 2006 – Accepted: 22 December 2006 – Published: 15 January 2007

First, we thank T. Schneider (TS hereafter) for his positive and constructive comments about Kondrashov and Ghil (2006) (KG hereafter). KG focused on exploiting temporal covariability in geophysical data sets, an idea that Schneider (2001; S01 hereafter) had suggested, but not applied to any data, synthetic or geophysical. Two unfortunate inaccuracies – corrected in comments (iii) and (iv) of TS – did crop up when KG described the expectation-maximization (EM) algorithm and its regularized version used by S01 for filling in missing data. We regret this slip, being thoroughly familiar with the general EM framework, which we used for probability density estimation when studying multiple weather regimes (Smyth et al., 1999; Kondrashov et al., 2004, 2006). We thus agree with TS that the regularized EM algorithm and KG's method are both based on estimating mean and covariance components of the gappy data set under study (his comment (iii)), and that several gap-filling methods, including regularized EM and our own (multi-channel) singular-spectrum analysis (M-)SSA, rely – among many other assumptions – also on the probability of a data point's absence being independent of the missing value itself (his comment (iv)).

Singular-value decomposition (SVD; Golub and Van Loan, 1989) underlies both regression and principal components analysis, and thus represents a common basis for KG's M-SSA as well as S01's regularized EM method. In this reply, we concentrate on discussing several differences between our methods, which might look minor to TS, but lead to differences in computational performance and numerical results in practical applications. We have tried out, before submitting KG, the free gap-filling software kindly provided on TS's personal website and plan to add KG's gap-filling feature as soon as feasible to the SSA-MTM Toolkit, available for free at http://www.atmos.ucla.edu/tcd/ssa; see also Ghil et al. (2002).

Correspondence to: D. Kondrashov
(dkondras@atmos.ucla.edu)

KG aim to fill the gaps with smooth information from an iteratively inferred "signal" that represents coherent spatio-temporal structures, and discard the "noise" variance. This idea is best illustrated by our synthetic example of a noise-contaminated oscillatory signal with a gap, see Fig. 2 in KG. Doing so can be quite valuable in a wide variety of applications, ranging from climate predictability and paleoclimate reconstruction to oceanographic and space physics data. To illustrate as simply as possible the difference between regularized EM (S01) and M-SSA (KG), consider a multivariate data set with just one missing value x_{ij} in one record at time t_j; the number of points i in each record (channels) is M, while the number of records (i.e., of sampling times t_j) is N. S01 (see Eq. 1 there) stresses regularization of the EM algorithm for rank-deficient cases, while in M-SSA regularization comes in the form of discarding "noise" EOFs, regardless of whether the data set is rank-deficient or not; see also Fig. A1 in Ghil et al. (2002).

For simplicity, we consider in this reply reconstruction based on spatial correlations only; our method is identical, in this case, to that of Beckers and Rixen (2003), while the emphasis in KG was on the use of purely temporal or mixed, spatio-temporal correlations. In our approach, we start an inner-loop iteration by computing the leading empirical orthogonal function (EOF) of the centered, zero-padded record. Then we perform the algorithm again on the new time series in which the principal component corresponding to that EOF alone was used to obtain nonzero values in place of the missing point and correct the channel's mean, the covariance matrix and the EOF. When this inner iteration has converged, we perform an outer-loop iteration by adding a second EOF for reconstruction and repeat the inner iteration. The outer iteration is performed only for *a few* significant, or "signal" EOFs, whose number is found by cross-validation. Beckers and Rixen (2003) discuss, in their Appendix A, how the bias introduced into the EOFs by missing data disappears as the iteration progresses.

Published by Copernicus GmbH on behalf of the European Geosciences Union and the American Geophysical Union.

In S01, iterations are also used in order to obtain an estimate of the missing value x_{ij}, along with the temporal mean $<x_i>$ in the spatial channel i, and the spatial covariance matrix. Estimating the regression coefficients in the EM algorithm with ridge regression has to be done record by record (see Eq. 2 in S01), including both "signal" and "noise" EOFs in the covariance matrix. As the number of records with missing data increases, KG offers potential computational savings since KG need (and want) to compute only a few leading EOFs in the outer iteration.

KG's double iteration differs, furthermore, from the Karhunen-Loève procedure for image processing of Everson and Sirovich (1995), where a previously fixed number of EOFs are estimated simultaneously in a single estimation loop. Everson and Sirovich's version of gap filling is closest to using the total truncated least-squares (TTLS) option in combination with the EM algorithm in S01. In this case, the relative computational performance of both KG and EM methods will largely depend on the number of EOFs involved, but ideas from both approaches could be useful in devising even better gap-filling methods in the future.

Edited by: B. D. Malamud
Reviewed by: T. Schneider and another referee

References

Beckers, J. and Rixen, M.: EOF calculations and data filling from incomplete oceanographic data sets, J. Atmos. Ocean. Technol., 20, 1839–1856, 2003.

Everson, R. M. and Sirovich, L.: The Karhuren-Loève transform for incomplete data, J. Opt. Soc. Am., 12, 1657–1664, 1995.

Ghil, M., Allen, R. M., Dettinger, M. D., Ide, K., Kondrashov, D., et al.: : Advanced spectral methods for climatic time series, Rev. Geophys., 40(1), 3.1–3.41, doi:10.1029/2000RG000092, 2002.

Golub, G. H. and Van Loan, C. F.: Matrix Computations (3rd ed.), The Johns Hopkins University Press, Baltimore and London, 728 pp, 1996.

Kondrashov, D., Ide, K., and Ghil, M.: Weather regimes and preferred transition paths in a three-level quasigeostrophic model, J. Atmos. Sci., 61, 568–587, 2004.

Kondrashov, D., Kravtsov, S., and Ghil, M.: Empirical mode reduction in a model of extratropical low-frequency variability, J. Atmos. Sci., 63, 1859–1877, 2006.

Smyth, P., Ide, K., and Ghil, M.: Multiple regimes in Northern Hemisphere height fields via mixture model clustering, J. Atmos. Sci., 56, 3704–3723, 1999.

Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, 2001.